Lead Score Case Study

Submitted by : Arpita Kachhwah

Lead Score Case Study for X Education

Problem Statement :

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Strategy

- Source the data for analysis
- Clean and prepare the data
- Exploratory Data Analysis.
- Feature Scaling
- Splitting the data into Test and Train dataset.
- Building a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

Problem solving methodology

Data Sourcing , Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis

Feature Standardization

- Feature Scaling and Splitting Train and Test Sets
- Feature Scaling of Numeric data
- Splitting data into train and test set.

Model Building

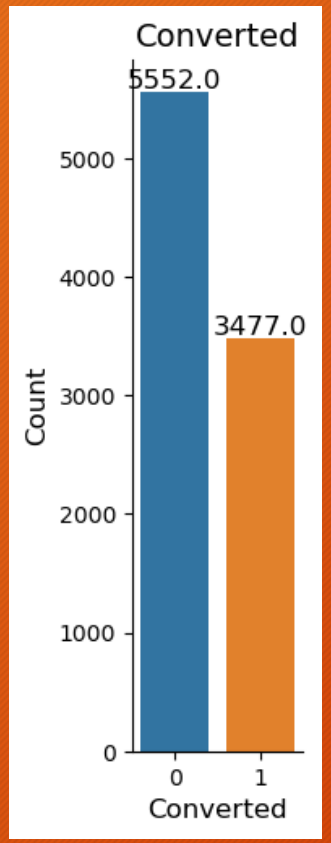
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

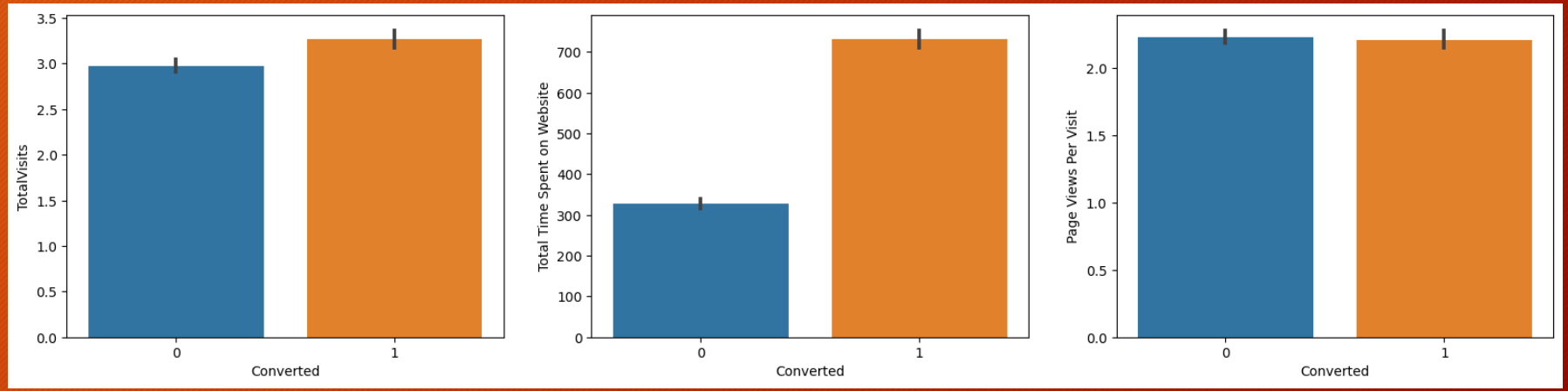
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

Exploratory Data Analysis

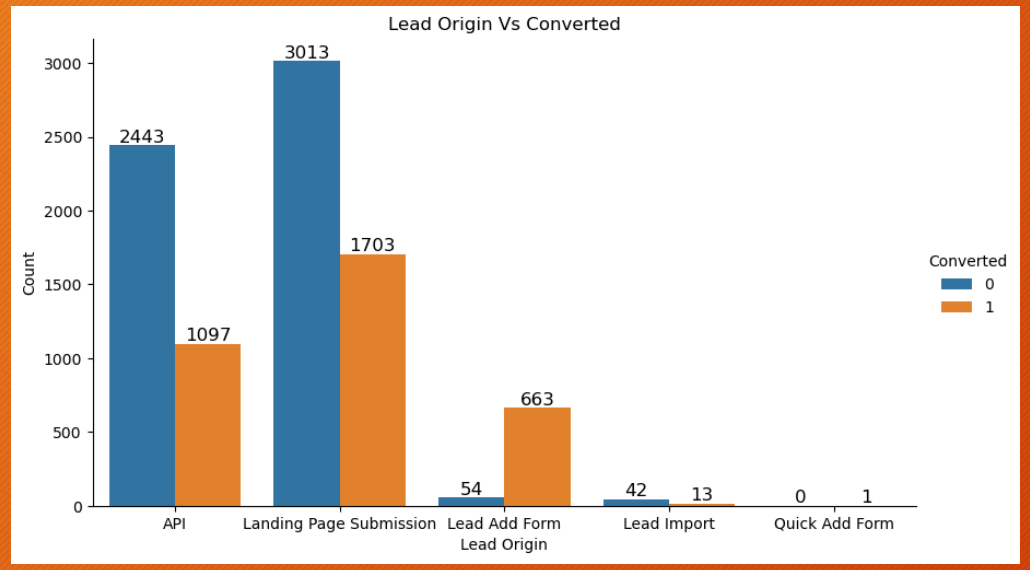
We have around 39% Conversion rate in Total



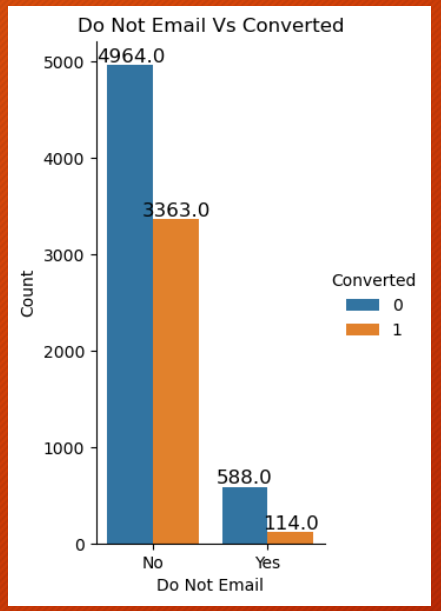
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

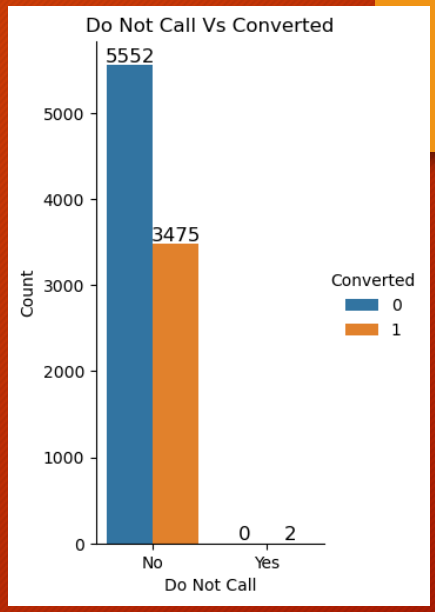


In Lead Origin, maximum conversion happened from Landing Page Submission

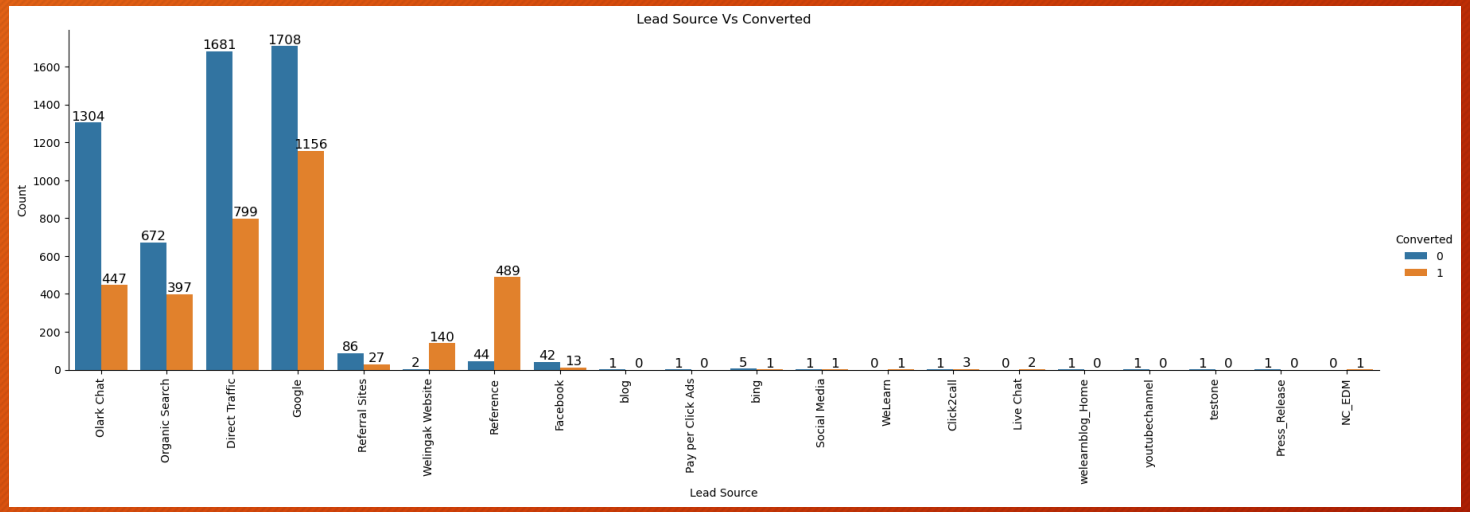


Major conversion has happened from Emails sent and Calls made

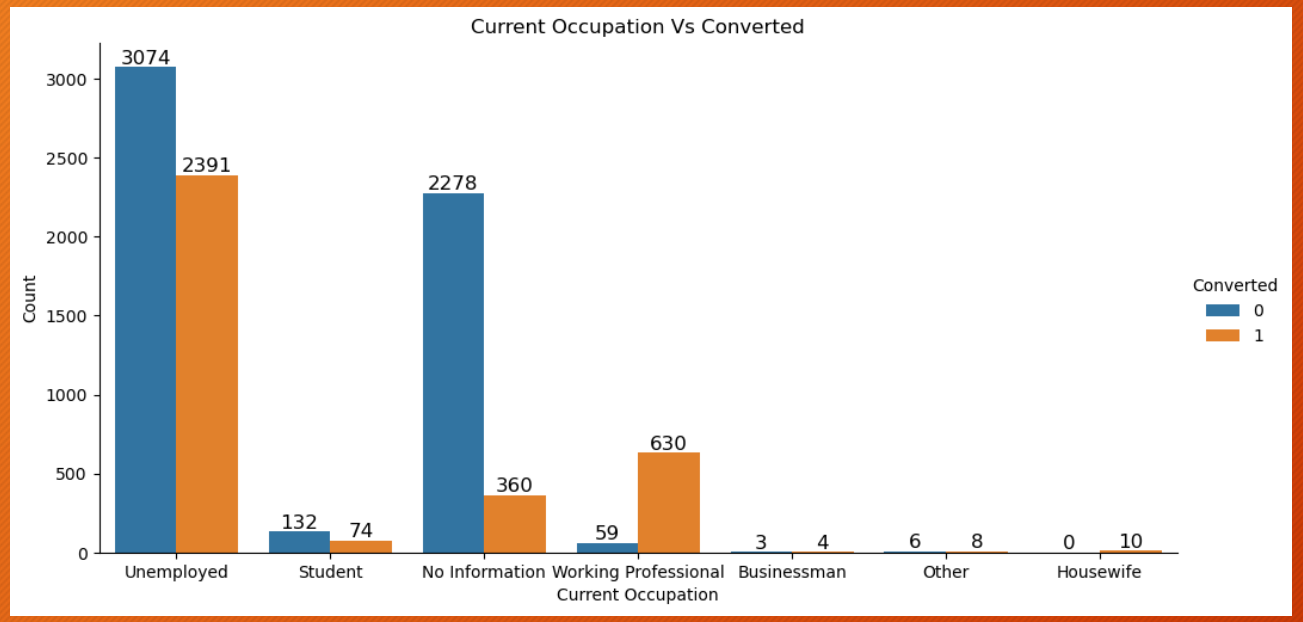




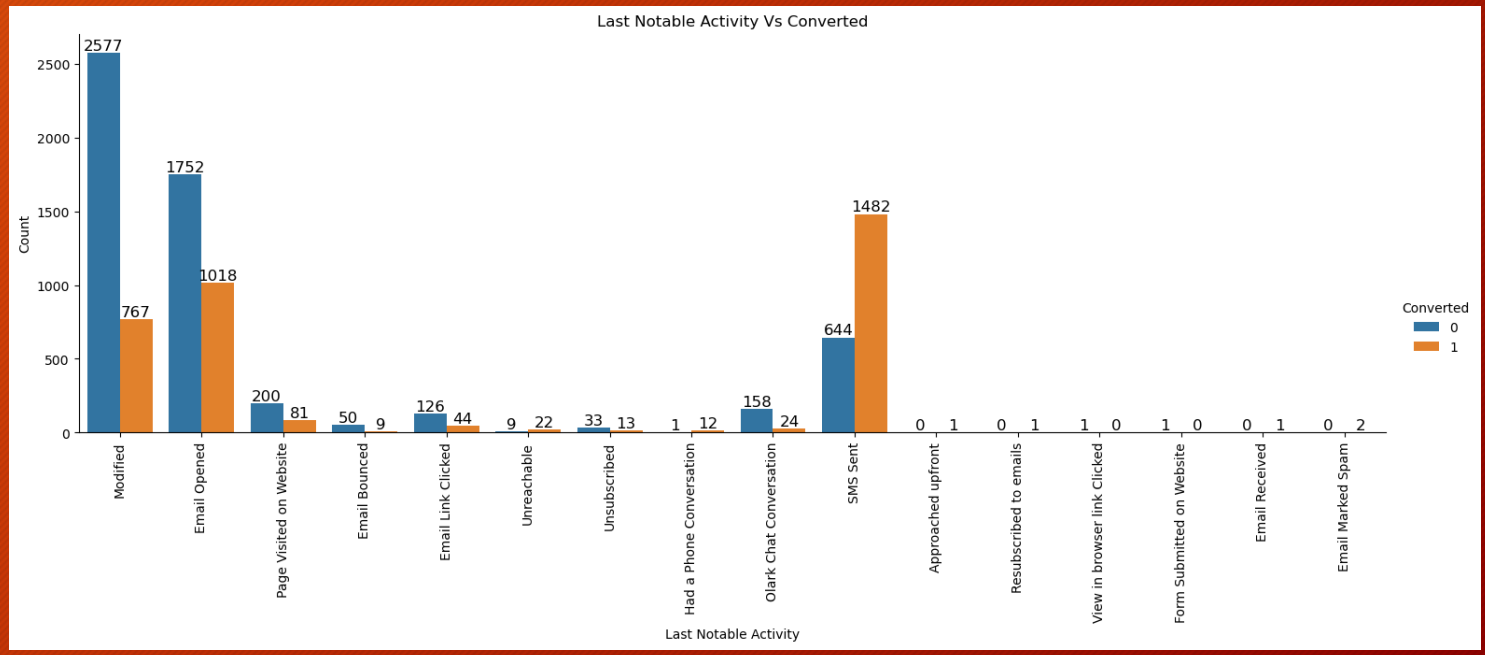
Major conversion in the lead source is from Google



More conversion happened with people who are unemployed



Last Activity value of SMS Sent' had more conversion.

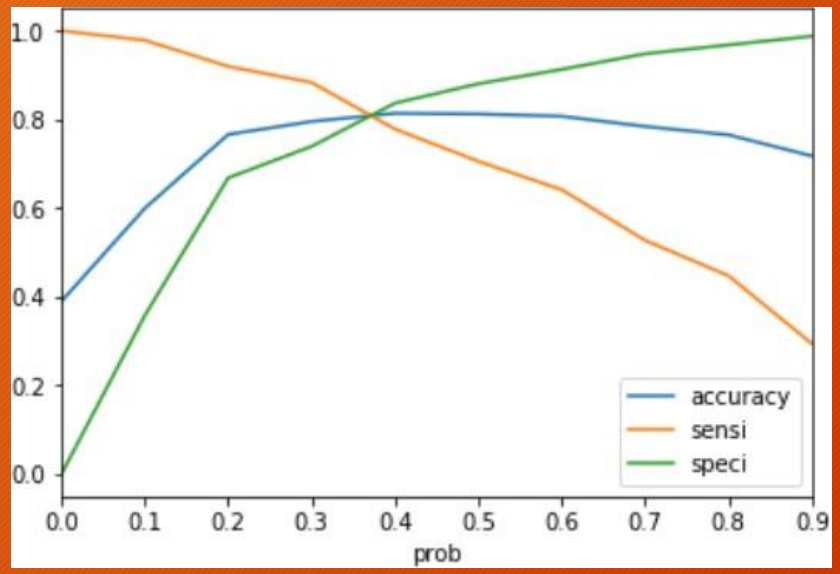


Variables Impacting the Conversion Rate

1. Do Not Email
2. Total Visits
3. Total Time Spent On Website
4. Lead Origin – Lead Page Submission
5. Lead Origin – Lead Add Form
6. Lead Source - Olark Chat
7. Last Source – Welingak Website
8. Last Activity – Email Bounced
9. Last Activity – Not Sure
10. Last Activity – Olark Chat Conversation
11. Last Activity – SMS Sent
12. Current Occupation – No Information
13. Current Occupation – Working Professional
14. Last Notable Activity – Had a Phone Conversation
15. Last Notable Activity - Unreachable

Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity



3161	697
974	1965

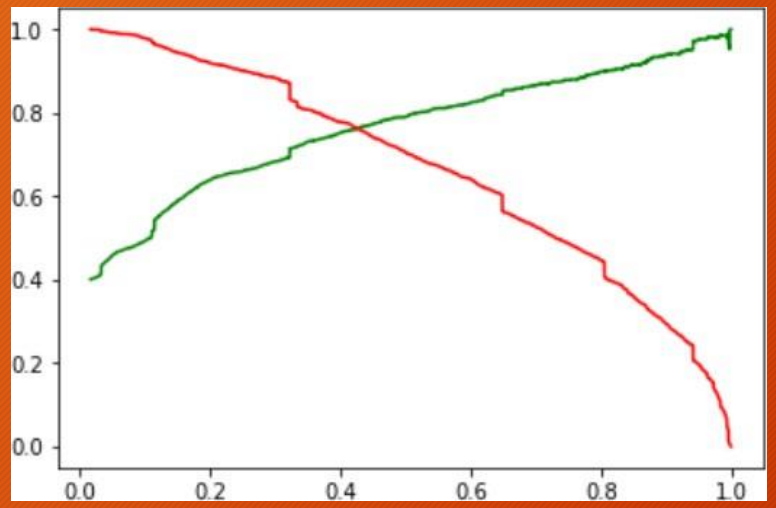
Confusion Matrix

- Accuracy - 81%
- Sensitivity - 80 %
- Specificity - 82 %
- False Positive Rate - 18 %
- Positive Predictive Value - 74 %
- Positive Predictive Value – 86%

Model Evaluation- Precision and Recall on Train Dataset

3397	461
725	1737

The graph depicts an optimal cut off of 0.42 based on Precision and Recall

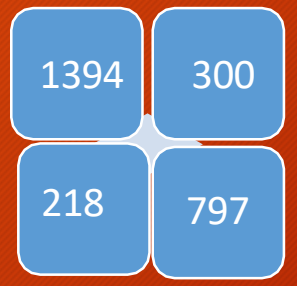


Confusion matrix

- Precision - 79 %
- Recall - 71 %

Model Evaluation – Sensitivity and Specificity on Test Dataset

- Accuracy - 81 %
- Sensitivity - 79 %
- Specificity - 82 %



Confusion Matrix

Conclusion

- We evaluated both Sensitivity-Specificity and Precision-Recall metrics to determine the optimal cut-off for the final prediction, focusing on Sensitivity and Specificity.
- The test set results show Accuracy, Sensitivity, and Specificity values of approximately 81%, 79%, and 82%, respectively, which are close to the values obtained from the training set.
- The calculated lead score indicates a conversion rate of around 80% in the training set and 79% in the test set.
- The top three variables contributing most to lead conversion in the model are:
  - Total Time Spent on Website
  - Lead Add Form (from Lead Origin)
  - Had a Phone Conversation (from Last Notable Activity)
- Overall, this model appears to be effective.